BIOMETRICS 55, 289–293
March 1999

Power and Sample Size for Testing Homogeneity of Relative Risks in Prospective Studies

Jun-Mo Nam

Biostatistics Branch, National Cancer Institute,
Executive Plaza North, Room 403, 6130 Executive Boulevard,
MSC 7368, Rockville, Maryland 20892-7368, U.S.A.
email: namj@epndce.nci.nih.gov

SUMMARY. Power and sample-size formulas for testing the homogeneity of relative risks using the score method are presented. The homogeneity score test (Gart, 1985, *Biometrika* **72**, 673–677) is formally equivalent to the Pearson chi-square test, although they look different. Results of this paper may be useful in assessing the validity of the model of a common relative risk before combining several 2×2 tables or in designing a prospective study for detecting heterogeneity of relative risks.

KEY WORDS: Power; Sample size; Score test; Testing homogeneity of relative risks; Prospective studies.

1. Introduction

The ratio of two binomial probabilities has been a parameter of major interest in biometry. For example, the ratio of proportions of having a disease for those exposed and those unexposed to a risk factor (i.e., relative risk or risk ratio) is a measure of the association between the disease and the risk factor in prospective studies. When more than one 2×2 table is involved in a study, we are also interested in comparing relative risks by strata. Gart (1985) has provided a likelihood score method for testing homogeneity of relative risks. The score test can be applied for detecting discrepancies in risk ratios or for examining the homogeneity assumption for a summary relative risk, depending on the nature of the study. Assuming the homogeneity, several authors (e.g., Gart, 1985; Gart and Nam, 1988) have suggested interval estimation of a common relative risk in the combination of the tables using likelihood scores. Their method possesses many desirable statistical properties. However, its use is not advisable when there are substantial discrepancies in risk ratios. It is prudent to examine the adequacy of the model of a common relative risk prior to its application.

Power and sample size related to the homogeneity score test have not been thoroughly studied. In Section 2 of this paper, we derive the asymptotic power of the score test and an approximate sample-size formula for prospective studies. Section 3 illustrates numerical examples using actual data, and Section 4 contains some concluding remarks. Note that we are interested in homogeneity of relative risks in cohort studies and not a common odds ratio in case–control studies.

2. Asymptotic Power and Sample Size

Consider J pairs of independent binomial variates, x_{0j} and x_{1j}, with corresponding parameters p_{0j} and p_{1j} with sample sizes n_{0j} and n_{1j} for $j = 1, 2, \ldots, J$. Let $q_{ij} = 1 - p_{ij}$ for $i = 0, 1$ and $j = 1, 2, \ldots, J$. Summation is denoted by dots, e.g., $x_{\cdot j} = x_{0j} + x_{1j}$ and $n_{\cdot j} = n_{0j} + n_{1j}$. Ratios of two binomial parameters are $p_{1j}/p_{0j} = \phi_j$ for $j = 1, 2, \ldots, J$. We are interested in testing the null hypothesis H_0: $\phi_j = \phi$ for every j against H_1: $\phi_j \neq \phi$ for any j. (Define $\boldsymbol{p}_0' = c p_{01}, p_{02}, \ldots, p_{0J}$).

The log-likelihood under the null hypothesis is

$$L_{\cdot}(\phi, \boldsymbol{p}_0) = \sum_j L_j(\phi, p_{0j}),$$

where $L_j(\phi, p_{0j}) = x_{1j}\ln(\phi) + (n_{1j} - x_{1j})\ln(1 - \phi p_{0j}) + x_{\cdot j}\ln(p_{0j}) + (n_{0j} - x_{0j})\ln(1 - p_{0j})$ for $j = 1, 2, \ldots, J$. The maximum likelihood estimators (MLEs) of ϕ and $\boldsymbol{p}_0$, $\hat{\phi}$ and $\hat{\boldsymbol{p}}_0$, are the solution of $J + 1$ equations from $\partial L_{\cdot}/\partial\phi = 0$ and $\partial L_j/\partial p_{0j} = 0$, i.e.,

$$\sum_j (x_{1j} - n_{1j}\hat{\phi}\hat{p}_{0j})/(1 - \hat{\phi}\hat{p}_{0j}) = 0$$

and

$$a_j \hat{p}_{0j}^2 + b_j \hat{p}_{0j} + c_j = 0,$$

where $a_j = n_{\cdot j}\hat{\phi}$, $b_j = -\{(x_{0j} + n_{1j})\hat{\phi} + x_{1j} + n_{0j}\}$, and $c_j = x_{\cdot j}$ for $j = 1, 2, \ldots, J$. Values of the MLEs are obtained by an iterative procedure (cf., Rao, 1965, pp. 302–305). The simple estimator of ϕ,

$$\hat{\phi}^{(0)} = \left\{\sum_{j=1}^{J} n_{0j}x_{1j}/(n_{\cdot j} - x_{\cdot j})\right\} \Big/ \left\{\sum_{j=1}^{J} n_{1j}x_{0j}/(n_{\cdot j} - x_{\cdot j})\right\}$$

(Tarone, 1981), may be used as an initial estimator for $\hat{\phi}$. The derivation of the simple estimator is analogous to that of the Mantel–Haenszel summary odds ratio (Mantel and Haenszel, 1959).

2.1 *Asymptotic Power*

Letting $S_{\phi j}(\phi, p_{0j}) = \partial L_j(\phi, p_{0j})/\partial\phi$, we can express a statistic for testing H_0 against H_1 as

$$X^2_{J-1} = \sum z_j^2(\hat{\phi}), \tag{1}$$

where

$$\begin{aligned} z_j(\hat{\phi}) &= S_{\phi j}(\hat{\phi}, \hat{p}_{0j}) / \left[\text{var}\left\{S_{\phi_j}(\hat{\phi}, \hat{p}_{0j})\right\}\right]^{\frac{1}{2}} \\ &= \left\{(x_{1j} - n_{1j}\hat{p}_{1j})/\hat{q}_{1j}\right\} / \left[n_{0j}n_{1j}\hat{p}_{1j}/\right. \\ &\quad \left.\left\{n_{1j}(\hat{\phi} - \hat{p}_{1j}) + n_{0j}\hat{q}_{1j}\right\}\right]^{\frac{1}{2}} \end{aligned}$$

for $j = 1, 2, \ldots, J$ (Gart, 1985). Note that $\hat{p}_{1j} = \hat{\phi}\hat{p}_{0j}$ for every j. Assume that n_{ij}'s are large and J is fixed. The statistic (1) is distributed asymptotically as a chi-square with $J-1$ degrees of freedom under H_0. The form (1) corresponds to the statistic for testing a common odds ratio in case–control studies (Breslow and Day, 1980), although they are different in explicit forms.

The score test statistic for homogeneity (1) is equivalent to the Pearson-type chi-square

$$X^2_{J-1} = \sum_{i=0}^{1}\sum_{j=1}^{J}(x_{ij} - n_{ij}\hat{p}_{ij})^2/(n_{ij}\hat{p}_{ij}\hat{q}_{ij}). \tag{2}$$

This is shown in Appendix 1.

Let $\chi^2_{J-1,(1-\alpha)}$ denote the $100 \times (1-\alpha)$ percentile point of a chi-square distribution with $J-1$ degrees of freedom. The asymptotic power of the homogeneity score test at level α is

$$\begin{aligned} &\Pr\left\{X^2_{J-1} \geq \chi^2_{J-1,(1-\alpha)} \mid H_1\right\} \\ &\quad = \Pr\left\{\chi^2_{J-1}(\Delta) \geq \chi^2_{J-1,(1-\alpha)}\right\}, \end{aligned} \tag{3}$$

where $\chi^2_{J-1}(\Delta)$ has a noncentral chi-square distribution with $J-1$ degrees of freedom and noncentrality parameter Δ, which is expressed as

$$\Delta = \sum n_{1j}(n_{1j}\phi\bar{q}_{0j} + n_{0j}\bar{q}_{1j})(p_{1j} - \bar{p}_{1j})^2/(n_{0j}\bar{p}_{1j}\bar{q}^2_{1j}),$$

where $\bar{p}_{1j} = \phi\bar{p}_{0j}$ for every j (A2.1). Using tables of the cumulative noncentral chi-square distribution (Haynam, Govindarajulu and Leone, 1970), we can obtain the power for given values of the noncentrality parameter and α level.

2.2 *Sample Size*

If the power is $1-\beta$, we have, from (3),

$$\chi^2_{J-1,\beta}(\Delta) = \chi^2_{J-1,(1-\alpha)}, \tag{4}$$

where $\chi^2_{J-1,\beta}(\Delta)$ is the $100 \times \beta$ percentile point of a noncentral chi-square distribution with $J-1$ degrees of freedom and noncentrality parameter Δ. Define design fractions as $t_j = n_{\cdot j}/N$, where $N = \Sigma\, n_{\cdot j}$ and $s_j = n_{1j}/n_{\cdot j}$, so that $n_{1j} = t_j s_j N$ and $n_{0j} = t_j(1-s_j)N$ for $j = 1, 2, \ldots, J$. We can express Δ in terms of N as

$$\Delta = \left\{\sum \frac{t_j s_j\left\{s_j\phi\bar{q}_{0j} + (1-s_j)\bar{q}_{1j}\right\}\left(p_{1j} - \bar{p}_{1j}\right)}{(1-s_j)\bar{p}_{1j}\bar{q}^2_{1j}}\right\} N,$$

where $\bar{p}_{1j} = \phi\bar{p}_{0j}$ for every j (A2.2). Note that ϕ and the $\bar{p}_{0j}$'s are found numerically by solving $J+1$ equations (A2.3). An analogous form of the noncentrality parameter is

$$\Delta = \left[\sum t_j\left\{\frac{s_j(p_{1j} - \bar{p}_{1j})^2}{\bar{p}_{1j}\bar{q}_{1j}} + \frac{(1-s_j)(p_{0j} - \bar{p}_{0j})^2}{\bar{p}_{0j}\bar{q}_{0j}}\right\}\right] N$$

(A2.4). The sample size required for a specific power of the score test for H_0 against H_1 at level α is found using the relation (4).

We calculated sample sizes required to achieve a specific power of the homogeneity score test for various values of relative risks, baseline probabilities, and design parameters. Some results for $J = 2$ are summarized in Table 1. Sample sizes based on simple and maximum likelihood estimates of ϕ, $N^{(0)}$, and N are essentially the same under a perfectly balanced design, i.e., $t_1 = t_2 = 1/2$ and $s_1 = s_2 = 1/2$. The sample size required is larger when baseline probabilities are smaller. It also relates inversely to the range of the variation among relative risks. Sample sizes calculated under other design configurations, e.g., $t_1 = 1/4$ and $t_2 = 3/4$ or $t_1 = 3/4$ and $t_2 = 1/4$, yield similar conclusions as above. It also demonstrates that the perfectly balanced design is optimal in terms of sample size. Approximate sample sizes calculated in Table 1 are based on asymptotic theory. We examine the accuracy of nominal powers for relatively small sample sizes by a Monte Carlo experiment using an IMSL subroutine (IMSL, 1987). Actual powers of the homogeneity score test for those total sample sizes less than or equal to 120 (i.e., $n = 30$) corresponding to a 50% nominal power in Table 1 (nine cases) range from 45 to 56% and those corresponding to an 80% power (two cases) are 78 to 79%, based on 1000 simulations. Note that the total sizes considered in simulations are slightly modified so that group size n is an integer. Also, undefined sampling points are excluded in computation of an actual power. The nominal powers are satisfactorily close to actual ones.

3. Numerical Examples

We illustrate application of the homogeneity score test, power, and sample-size calculations in the following two examples.

Example 1

In a carcinogenesis bioassay study of Avadex (Innes et al., 1969), the fungicide was orally administered to both males (M) and females (F) in two strains of mice (X and Y). Frequencies of pulmonary tumors among test mice for categories XM, XF, YM, and YF were 4/16, 2/16, 4/18, and 1/15, and their respective controls were 5/79, 3/87, 10/90, and 3/82. Corresponding relative risks were 3.95, 3.63, 2.00, and 1.82. The score test (1) does not reject homogeneity of relative risks ($X^2_3 = 0.954$, $p = 0.81$), and it enables us to combine information on a common relative risk from the four 2×2 tables. The 95% confidence interval for a common relative risk (Gart and Nam, 1988) is (1.35, 5.03), and it is concluded that the fungicide is tumorigenic in mice. We examine the asymptotic power of the homogeneity test and approximate sample-size requirement for this experiment. The relative risks for those animals exposed to the fungicide are $\phi_1 = 3.95$, $\phi_2 = 3.63$, $\phi_3 = 2.00$, and $\phi_4 = 1.82$. Under the perfectly balanced design ($t_j = 0.25$ and $s_j = 0.5$ for every j), we

Table 1

Approximate sample sizes required for 50 and 80% power of the homogeneity score test at $\alpha = 0.05$ under the perfectly balanced design with $J = 2$ (N and $N^{(0)}$ are sample sizes based on simple and maximum likelihood estimates of a common relative risk)

Baseline probabilities		Relative risks		50% power		80% power	
p_{01}	p_{02}	ϕ_1	ϕ_2	$N^{(0)}$	N	$N^{(0)}$	N
0.1	0.05	0.5	1.5	732	732	1496	1496
			3.0	231	231	472	472
			5.0	122	120	249	245
		1.0	2.0	1449	1449	2961	2961
			3.0	522	522	1067	1067
			5.0	215	213	439	435
		2.0	5.0	614	614	1254	1254
			8.0	239	232	488	474
0.2	0.10	0.5	1.5	339	339	693	693
			3.0	105	105	214	214
			5.0	54	51	110	105
		1.0	2.0	659	659	1346	1346
			3.0	234	235	478	480
			5.0	94	92	192	188
		2.0	5.0	268	270	548	551
			8.0	100	87	205	177

obtain $\Delta = (0.003675)N$ from (A2.2). From (3), the approximate power of the homogeneity test of size $\alpha = 0.05$ is 15.4% for $N = 411$. The magnitude of the variation among these relative risks is not large enough to be detected by the homogeneity test with a reasonable power.

Unless the degree of variation is very large, the assumption of homogeneity is not likely to be rejected in a routine carcinogenesis screening experiment. The noncentrality parameter corresponding to 50% power of the noncentral chi-square with 3 d.f. at $\alpha = 0.05$ is 5.760; we then have $N = 5.760/0.003675 = 1567$. It is clearly a size beyond the limitations of animal experiments. Note that homogeneity of relative risks is not the main concern in this carcinogenesis study. Rather, it serves as a rational basis for combining several 2×2 tables in a common relative risk.

Example 2

Evans et al. (1978) investigated the relation between oral contraceptive (OC) use and bacteriuria in a population-based cohort of women aged less than 50 years. The total sample size was 2357. The rates of women with bacteriuria among OC users and nonusers were 23/441 and 65/1456 for ages 16–39 years and 4/18 and 12/452 for ages 40–49 years. The relative risk for the first age group was $\phi_1 = 1.11$ while that for the second group was $\phi_2 = 8.37$. The latter was eightfold greater than the former. The study indicated no association between OC use and bacteriuria for the younger group but a very strong positive association for the older group. Conflicting reports in prior studies (e.g., Kunin and McCormack, 1968; Sussman et al., 1969; Takahashi and Loveland, 1974) could be explained by the age factor. From (1), the score test for detecting heterogeneity of two relative risks is highly significant ($X_1^2 = 14.55$, $p = 0.0001$). From (3) and (A2.1), we obtain the power of this test as 97%. From (4) and (A2.2), we have a sample size required for power = 80% of the score test at the 0.05 level as $N = 1214$ for $p_{01} = 0.045$, $p_{02} = 0.027$, $\phi_1 = 1.11$, $\phi_2 = 8.37$, $s_1 = 0.25$, $s_2 = 0.4$, $t_1 = 0.8$, and $t_2 = 0.2$. Evans et al. need only half their sample to detect heterogeneity of these relative risks with good power.

4. Remarks

In stratified cohort studies, researchers intend to find an association of a major risk factor with disease. The benefit of combining relative risks on the summary risk is quite striking in terms of the width of a confidence interval for small or medium sample sizes. Detecting a small variation among risk ratios with good power requires a very large sample size. Efficient interval estimation of a common relative risk and the optimum score test (Gart and Nam, 1988) can be justifiably applied in typical carcinogenesis bioassay experiments (e.g., Sontag, Page, and Saffiotti, 1976). In some cohort studies, relative risks vary with the strata and researchers aim to detect heterogeneity among relative risks. Formulas of power and sample size for the score test can be helpful in designing such prospective studies. Jones et al. (1989) have empirically examined homogeneity tests of the odds ratio across strata in case–control studies and have reported low power, which parallels the case of the common relative risk in cohort studies.

In comparative prospective studies, the relative risk is the measure for appraising a causal effect of an agent to a particular disease and the odds ratio is not of interest per se (e.g., Miettinen, 1985). For estimating a common relative risk across strata in prospective studies, e.g., example 1 (carcinogenesis bioassay experiment) in Section 3, the validity of the homogeneity of relative risks can be examined by the likelihood score test (Gart, 1985) but not usually by a test for a common odds ratio (e.g., Breslow and Day, 1980). The former is specifically designed for prospective studies, while the latter is for case–control studies. In definition, relative

risk (ratio of two incidence rates) and odds ratio (ratio of two odds) are different parameters. In this paper, we have limited our investigation to homogeneity of relative risks for stratified prospective studies.

Radhakrishna (1966) generalized the approach of Cochran (1954) to testing the equality of response measures of two treatments across strata assuming a constant difference between the measures on various scales. Note that common relative risk (ϕ), odds ratio (ψ), and difference (δ) are response measures with a constant difference on the logarithmic, logit, and constant scales, respectively. When sample sizes are relatively small, the homogeneity models are not likely to be rejected. If the goal of the analysis is testing rather than estimation, the maximin efficient robust test for equality of proportions across strata on a family of scales (Gastwirth, 1985), which attains a relatively high efficiency, may be prudent. Applied to the bioassay data (Example 1, Section 3), the maximin test for detecting a tumorigenicity of the fungicide is highly significant ($p = 0.004$), with asymptotic relative efficiency (ARE) 93%. In stratified comparative studies, however, researchers are generally not only interested in hypothesis testing but also in estimation, assuming homogeneity of the parameter of interest across strata. The choice of parameter to be estimated (e.g., relative risk, odds ratio) may be based on the nature of the study, biological justification, common practice, and/or empirical preference.

Acknowledgements

The author wishes to thank the referees and associate editor for helpful comments and Mrs Jennifer Donaldson for the careful typing of the manuscript.

Résumé

Les formules de calcul de puissance et de détermination de tailles d'échantillons sont présentées pour le test de l'homogénéité des risques relatifs à l'aide de la méthode du score. Le score-test d'homogénéité (Gart, 1985, *Biometrika* **72,** 673–677) est formellement équivalent au test de chi-deux de Pearson bien que présenté différemment. Les résultats de cet article peuvent être utiles pour évaluer la validité d'un modèle de risque relatif commun avant de combiner plusieurs tables 2 × 2 ou dans la conception d'une étude prospective pour détecter l'hétérogénéité de risques relatifs.

References

Breslow, N. E. and Day, N. E. (1980). *Statistical Methods in Cancer Research,* Volume 1, *The Analysis of Case-Control Studies.* Lyon: International Agency for Research on Cancer.

Cochran, W. G. (1954). Some methods for strengthening the common χ^2 tests. *Biometrics* **10,** 417–451.

Evans, D. A., Hennekens, C. H., Miao, L., Laughlin, L. W., Chapman, W. G., Rosner, B., Taylor, J. O., and Kass, E. H. (1978). Oral-contraceptive use and bacteriuria in a community-based study. *New England Journal of Medicine* **299,** 536–537.

Gart, J. J. (1985). Approximate tests and interval estimation of the common relative risk in the combination of 2 × 2 tables. *Biometrika* **72,** 673–677.

Gart, J. J. and Nam, J.-M. (1988). Approximate interval estimation of the ratio of binomial parameters: A review and corrections for skewness. *Biometrics* **44,** 323–338.

Gastwirth, J. L. (1985). The use of maximin efficiency robust tests in combining contingency tables and survival analysis. *Journal of the American Statistical Association* **80,** 380–384.

Haynam, G. E., Govindarajulu, J., and Leone, F. C. (1970). Tables of the cumulative non-central chi-square distribution. In *Selected Tables in Mathematical Statistics,* Volume I, H. L. Harter and D. B. Owen (eds), 1–78. Chicago: Markham.

IMSL. (1987). *IMSL User's Manual: Stat/Library (Fortran Subroutines for Statistical Analysis),* Version 1.0. Houston: IMSL.

Innes, J. R. M., Ulland, B. M., Valerio, M. G., Petrucelli, L., Fishbein, L., Hart, E. R., and Pallotta, A. J. (1969). Bioassay of pesticides and industrial chemicals for tumorigenicity in mice: A preliminary note. *Journal of the National Cancer Institute* **42,** 1101–1114.

Jones, M. P., O'Gorman, T. W., Lemke, J. H., and Woolson, R. F. (1989). A Monte Carlo investigation of homogeneity tests of the odds ratio under various sample size configurations. *Biometrics* **45,** 171–181.

Kunin, C. M. and McCormack, R. C. (1968). An epidemiologic study of bacteriuria and blood pressure among nuns and working women. *New England Journal of Medicine* **278,** 635–642.

Mantel, N. and Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute* **22,** 719–748.

Miettinen, O. S. (1985). *Theoretical Epidemiology.* New York: Wiley.

Radhakrishna, S. (1965). Combination of results from several 2 × 2 contingency tables. *Biometrics* **21,** 86–98.

Rao, C. R. (1965). *Linear Statistical Inference and Its Application.* New York: John Wiley.

Sontag, J. A., Page, N. P., and Saffiotti, U. (1976). *Guidelines for Carcinogen Bioassay in Small Rodents.* NIH Publication 76-801. Bethesda, Maryland: National Institutes of Health.

Sussman, M., Asscher, A. W., Waters, W. E., Evans, J. A., Campbell, H., Evans, K. T., and Williams, J. E. (1969). Asymptotic significant bacteriuria in the non-pregnant women. I: Description of a population. *British Medical Journal* **1,** 799–803.

Takahashi, M. and Loveland, D. B. (1974). Bacteriuria and oral contraceptives: Routine health examinations of 12,076 middle-class women. *Journal of the American Medical Association* **227,** 762–765.

Tarone, R. E. (1981). On summary estimators of relative risk. *Journal of Chronic Disease* **34,** 463–468.

Received May 1997. *Revised October* 1997 *and March* 1998. *Accepted April* 1998.

Appendix 1

Equivalence of the Score and Pearson's Chi-Square Tests

From (1), the jth component of the likelihood score statistic is written as

$$\begin{aligned}&\{(x_{1j}-n_{1j}\hat{p}_{1j})/\hat{q}_{1j}\}^2/[n_{0j}n_{1j}\hat{p}_{1j}/\{n_{1j}(\hat{\phi}-\hat{p}_{1j})+n_{0j}\hat{q}_{1j}\}]\\&\quad=\{(x_{1j}-n_{1j}\hat{p}_{1j})^2/(n_{1j}\hat{p}_{1j}\hat{q}_{1j})\}\{1+n_{1j}\hat{\phi}\hat{q}_{0j}/(n_{0j}\hat{q}_{1j})\}\\&\quad=(x_{1j}-n_{1j}\hat{p}_{1j})^2/(n_{1j}\hat{p}_{1j}\hat{q}_{1j})\\&\qquad+(x_{1j}-n_{1j}\hat{p}_{1j})^2\hat{\phi}\hat{q}_{0j}/(n_{0j}\hat{p}_{1j}\hat{q}_{1j}^2).\end{aligned}\tag{A1.1}$$

Using the relation,

$$(x_{1j}-n_{1j}\hat{p}_{1j})/\hat{q}_{1j}=-(x_{0j}-n_{0j}\hat{p}_{0j})/\hat{q}_{0j},\tag{A1.2}$$

from $(\partial L_j/\partial p_{0j})_{p_{0j}=\hat{p}_{0j},\phi=\hat{\phi}}=0$, the jth component can be expressed as

$$\sum_{i=0}^{1}(x_{ij}-n_{ij}\hat{p}_{ij})^2/(n_{ij}\hat{p}_{ij}\hat{q}_{ij}).$$

The summation over j leads to (1) = (2).

Appendix 2

Derivation of Noncentrality Parameter

Under H_1: $\phi_j=\phi$ for any j, the statistic (1) is an asymptotically noncentral chi-square with $J-1$ degrees of freedom and a noncentrality parameter

$$\Delta=\sum n_{1j}(n_{1j}\phi\bar{q}_{0j}+n_{0j}\bar{q}_{1j})(p_{1j}-\bar{p}_{1j})^2/(n_{0j}\bar{p}_{1j}\bar{q}_{1j}^2),\tag{A2.1}$$

where $\bar{p}_{1j}=\phi\bar{p}_{0j}$ for every j. The $\bar{p}_{0j}$ is the asymptotic value of $\hat{p}_{0j}$ under H_0 and the solution of the quadratic equation $\bar{a}_j\bar{p}_{0j}^2+\bar{b}_j\bar{p}_{0j}+\bar{c}_j=0$, where $\bar{a}_j=n_{\cdot j}\phi$, $\bar{b}_j=-\{(n_{0j}p_{0j}+n_{1j})\phi+n_{1j}p_{1j}+n_{0j}\}$, and $\bar{c}_j=n_{0j}p_{0j}+n_{1j}p_{1j}$. We obtain (A2.1) from the asymptotic expectation of $z_j^2(\hat{\phi})$ for every j under H_1.

Since $n_{1j}=t_js_jN$ and $n_{0j}=t_j(1-s_j)N$, where $N=\Sigma n_{\cdot j}$, (A2.1) is rewritten as

$$\Delta=\left\{\sum\frac{t_js_j\{s_j\phi\bar{q}_{0j}+(1-s_j)\bar{q}_{1j}\}(p_{1j}-\bar{p}_{1j})^2}{(1-s_j)\bar{p}_{1j}\bar{q}_{1j}^2}\right\}N,\tag{A2.2}$$

where $\bar{p}_{1j}=\phi\bar{p}_{0j}$ for every j. Note that ϕ and the $\bar{p}_{0j}$'s are numerically found by solving the following $J+1$ equations:

$$\sum_{j=1}^{J}t_js_j(p_{1j}-\phi\bar{p}_{0j})/(1-\phi\bar{p}_{0j})=0$$

and

$$\bar{p}_{0j}=-\left\{\bar{b}_j+(\bar{b}_j^2-4\bar{a}_j\bar{c}_j)^{\frac{1}{2}}\right\}/(2\bar{a}_j),\tag{A2.3}$$

where $\bar{a}_j/N=t_j\phi$, $\bar{b}_j/N=-t_j\{(p_{0j}+s_jq_{0j})\phi+(1-s_jq_{1j})\}$, and $\bar{c}_j/N=t_j\{(1-s_j)p_{0j}+s_jp_{1j}\}$ for $j=1,2,\ldots,J$. An initial value for ϕ is

$$\phi^{(0)}=\sum_{j=1}^{J}\frac{t_js_j(1-s_j)p_{1j}}{(1-s_j)q_{0j}+s_jq_{1j}}\Bigg/\sum_{j=1}^{J}\frac{t_js_j(1-s_j)p_{0j}}{(1-s_j)q_{0j}+s_jq_{1j}}.$$

The noncentrality parameter of the asymptotic power function of the Pearson-type chi-square test is expressed as

$$\Delta=\left[\sum t_j\left\{\frac{s_j(p_{1j}-\bar{p}_{1j})^2}{\bar{p}_{1j}\bar{q}_{1j}}+\frac{(1-s_j)(p_{0j}-\bar{p}_{0j})^2}{\bar{p}_{0j}\bar{q}_{0j}}\right\}\right].\tag{A2.4}$$

Using the relation (A1.2) under H_1 as $N\to\infty$,

$$n_{1j}(p_{1j}-\bar{p}_{1j})/\bar{q}_{1j}=-n_{0j}(p_{0j}-\bar{p}_{0j})/\bar{q}_{0j},$$

we show that the noncentrality parameter (A2.1) is equal to (A2.4). Since the statistics (1) and (2) are equivalent (Appendix 1), their powers should be the same.